論文 / 著書情報
Article / Book Information

Title	Visualization of Audio Information for Home Video Highlight Extraction
Author	Koichi Takagi, Ryoichi Kawada, Takahiro Shinozaki, Sadaoki Furui
Journal/Book name	Proc. of the Second APSIPA Annual Summit and Conference, , , pp. 145-148
Issue date	2010, 12

Visualization of Audio Information for Home Video Highlight Extraction

Koichi Takagi*†, Ryoichi Kawada*, Takahiro Shinozaki† and Sadaoki Furui†
*KDDI R&D Laboratories Inc., Fujimino, Saitama 356-8502 Japan
E-mail: ko-takagi@kddilabs.jp Tel: +81-49-278-7432
†Tokyo Institute of Technology, Ookayama, Tokyo 152-8550 Japan

Abstract—**This paper proposes a method for supporting highlight extraction from a home video on a mobile terminal using both audio and visual information but without having to directly listening to the audio. This study makes two main contributions. First, we analyze the difference in the highlight extraction results with or without listening to the audio and then identify the important audio information, which cannot be obtained simply by watching the video. Second, based on the results of the analysis, essential audio information is extracted and visualized on the small display of a mobile terminal. The experimental results show that the effectiveness of the visualized audio information during highlight extraction from home video is comparable to that obtained by listening to the audio.**

I. Introduction

In recent years, there is no problem in making a long movie (home video) because the storage space on the mobile terminals has become larger. However, when shared with other people, the video needs to be edited to the proper length for viewing. Therefore, there is a demand to easily extract only the important parts (highlights) from the original home video. Moreover, highlights need to be easily extracted anytime, anywhere, as shown in Fig. 1.

Fig. 1 People operating their mobile terminals anywhere (from the left, on a train, in a restaurant, and in university free space.)

Most home videos include both *video* and *audio* information (hereafter, denoted as V and A, respectively). When highlights are extracted from home video, not only V data but also A data are generally referred. Moreover, in order to easily execute this task on a mobile terminal, it is desirable for it to be done simply by using the displayed data without taking the time to listen to the audio information.

For this reason, there have been many studies on audio indexing, classification, and visualization. For example, in order to author broadcast data, each A segment is annotated by analyzing A information [1]. After speech and music intervals from broadcast data are respectively detected, they are indexed as a part of video data. Basically, A information is treated independently of V information. On the other hand, many conventional studies index video segments by means of both V and A data. (There are some surveys [2][3] regarding video skimming or video summarization, which are generally carried out after highlight extraction.) In these studies, both A and V information are treated at the same level. There is no research on audio information visualization considering ways to view important audio information together with video without listening, in the application of manual highlight extraction.

We focus on an appropriate visualization method of audio data under conditions where only visualized data can be observed. In order to find the answer to this requirement, we first investigate the difference between the cases where (a) both V and A are presented to subjects (hereafter denoted as "V+A") and where (b) only V is presented to the subjects. We analyze and identify the essential A information that causes the differences. Based on these results, we propose a method for extracting and visualizing essential A information (hereafter denoted as "*visA*"). In order to verify the proposed method, we make a comparison of V+A and V+*visA* cases. Finally, we conclude this paper.

II. Video-Audio Relation on Highlight Extraction from a Home Video

The requirement for this research is to ascertain useful and essential A information, which cannot be obtained from V, for the highlight extraction task. For example, A information that cannot be obtained from V data alone is the sound coming from objects that do not appear in the video. However, such sound data are not always needed. In this section, the difference between the results of highlight extraction from *Vonly* and V+A cases is investigated.

A. Preliminary experiments

TABLE I
Experimental Setup (two datasets differ from each other.)

	Section II	Section IV
# of total sequences	40 (= about an hour)	40 (= about an hour)
# of presentation sequences per subject	5 sequences x 2 (*Vonly*, V+A)	5 sequences x 3 (*Vonly*, V+*visA*, V+A)
# of subjects	16 (twenties-forties, male and female)	16 (twenties-forties, male and female)
Specification of sequences	Captured by mobile terminals: Original duration: 30sec. to 2 min. File format: MP4 format(3g2/3gp) Video: H.264, 15fps, QCIF-QVGA, 64-256kbps Audio: AMR-NB (8kHz, mono, 12.2kbps)	
Contents of sequences	Home videos captured by mobile terminals For example, taking a walk around a park, eating dinner at a dining room, playing with dogs, driving, shopping, sightseeing, etc.	

A subjective evaluation of highlight extraction is carried out in this section. Experimental conditions are shown in TABLE I. Each subject is asked to "extract highlights from the video

Proceedings of the Second APSIPA Annual Summit and Conference, pages 145–148,
Biopolis, Singapore, 14-17 December 2010.

so that the total duration is 20%-30% of the original sequence," and then evaluates N kinds of sequences. Highlights are extracted by each subject based on the presentation of either *Vonly* or *V+A* for each sequence. In other words, highlights are extracted twice for each sequence. If subjects evaluate the *Vonly* case after the *V+A* case, they may process the *Vonly* case based on the *A* information that they can recall. Moreover, two consecutive sequences need to be different so as not to bore the subjects. Therefore, we decided to let each subject finish all the sequences by *Vonly*, and then process the same sequences again by *V+A*.

B. *Results and considerations*

In the results, there is a clear difference between Vonly and V+A conditions, as we expected. The results are shown in Fig. 2. First, the audio contents (which are allowed to overlap each other) for each segment (1 second) are labeled manually. The contents are labeled by one of the audio indices in TABLE II. Then, the extracted highlight segment by *V+A* for each subject/sequence is treated as the ground truth and compared with that by *Vonly*. The effectiveness of audio information is evaluated by counting the number of commonly extracted (Co-corr.) 1-s periods, and inserted (Ins.) or deleted (Del.) periods under the *Vonly* condition in comparison with the *V+A* condition. These results are also shown in Table II. The "# of segments" in this table denotes the number of total periods corresponding to the presented sequences.

As shown in this table, more "human voice" segments are extracted in the *V+A* case compared with the *Vonly* case. Many users tend to extract the "human voice" segment by listening to the audio.

The "sudden sound," such as "machine noise" and "impulse noise," also shows a large difference between the *Vonly* and *V+A* counts. The segments of "sudden sound" are both inserted and deleted as highlight segments.

On the other hand, the number of "cheering" and "applause" segments, which are generally extracted, shows little change. The reason is that (a) these segments can be recognized as highlights from *Vonly*, and (b) these segments are not extracted as highlights for home video compared with broadcast video. In any case, these segments are not necessary for our purpose.

In summary, the number of extracted segments for both "human voice" and "sudden sound" shows large changes. Therefore, if we can extract such segments and display the information on a timeline, our purpose can be achieved.

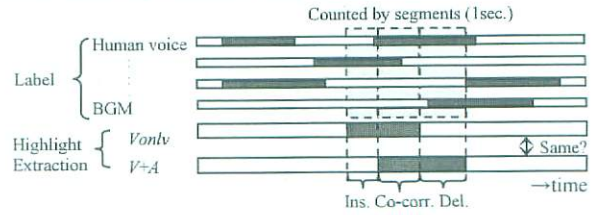

Fig. 2 Differences of extracted highlight segments under *Vonly* or *V+A* conditions.

III. AN AUDIO VISUALIZATION METHOD

This section describes how the essential information is extracted from audio data based on the results of the previous section and how the data are presented. The overview is shown in Fig. 3.

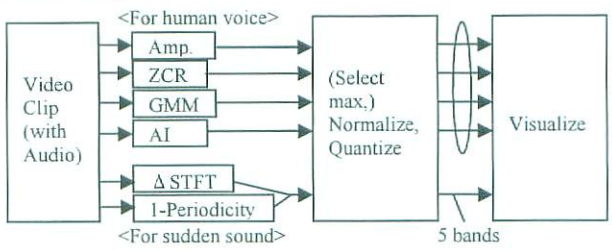

Fig. 3 An overview of the audio visualization scheme.

A. *Extraction of essential audio information*

The necessary audio information is essentially the "human voice" and "sudden sound." The followings describe the detection methods.

(a) Human voice

There are many conventional research studies on the detection of the human voice, which is generally called VAD (Voice Activity Detection). However, it is not preferable for detecting all kinds of human voices. Therefore, we should choose the "human voice" that is required for this purpose. As the result of observations on preliminarily experiments, extracted human voice characteristics as highlights are *loud* and *distinct*. Conversely, the human voice, where the level is low or drowned out by other loud sounds, is not often extracted as highlights. In view of these facts, the following

TABLE II
MATCHING ACCURACY OF HIGHLIGHT EXTRACTION BETWEEN *VONLY* AND *V+A*.

Index	Description	# of segments	Del.	Ins.	Co-corr.
Human voice	Human voice	1316	**234(17.8%)**	62(4.7%)	**259(19.7%)**
Machine noise (Bell,etc.)	Sound produced by machines (such as a bell)	248	**22(8.9%)**	**32(12.9%)**	51(20.6%)
Vehicle (w/o bell)	Sound from vehicles	315	10(3.2%)	12(3.8%)	52(16.5%)
Animal	Sound from animals	143	8(5.6%)	2(1.4%)	13(9.1%)
Cheering	A shout of encouragement, approval, etc.	621	21(3.4%)	23(3.7%)	101(16.3%)
Applause	Hand clapping	321	18(5.6%)	12(3.7%)	71(22.1%)
BGM	Back ground music	1437	65(4.5%)	105(7.3%)	72(5.0%)
Musical instrument (w/o BGM)	Sound from musical instrument	312	12(3.8%)	2(0.6%)	23(7.4%)
Water, wave	Sound regarding water	102	2(2.0%)	0(0.0%)	9(8.8%)
Wind	Sound from wind	82	0(0.0%)	3(3.7%)	8(9.8%)
Life sounds	Abuzzing sound , footsteps, etc.	1050	13(1.2%)	12(1.1%)	35(3.3%)
Impulse sound	Crash, knocking, etc.	612	**39(6.4%)**	**41(6.7%)**	42(6.9%)

three methods are used [4] .[1]

- *Amplitude level*

The amplitude level is one of the conventional features for VAD, though it is not robust under noisy conditions. It is defined as the logarithm of the signal energy; that is, for each N-sample Hamming-windowed frame $\{x_n: n=1,\ldots,N\}$, it is computed as

$$(1/N)\sum_n \log(x_n)\cdot \quad (1)$$

- *Zero crossing rate (ZCR)*

The zero crossing rate (ZCR) is the number of times the signal level crosses 0 during each frame. It is very effective for some kinds of noise, but not at all for noise with frequent zero crossings.

- *Speech/Non-speech GMM likelihood*

The Gaussian mixture model (GMM) is widely used for speech detection, because the statistical model is easily trained and usually effective. The log-likelihood ratio of speech GMM to non-speech GMM for an input frame is computed by

$$\log(p(v_t \mid \Theta_s)) - \log(p(v_t \mid \Theta_n)), \quad (2)$$

where v_t is an acoustic vector for the GMMs, and Θ_s and Θ_n denote the model parameter set for speech and noise, respectively. Each GMM consists of 16 Gaussians with diagonal covariance matrices, and a 38-dimensional feature vector (12 MFCCs, their first/second derivatives, Δ power and $\Delta\Delta$ power) is used. The speech/non-speech GMM is trained with the labeled sequence used in section II.

The above-mentioned features do not consider whether each utterance is distinct. Therefore, we adopt a voice articulation index (AI) that denotes a voice clarity feature in addition to the above-listed features.

- *AI*

The count-the-dot method [4] is used for simplicity. These dots are determined for every frequency and only counted according to the power spectrum.

The frame length is 25 msec for the GMM likelihood and 100 msec for the other measurements. The frame shift is 50% for all features. Each feature value of a segment is represented by selecting the maximum frame feature value. In order to normalize these values x from 0 to 1, we use a sigmoid function

$$f(x) = \left[1+\exp\left\{-\frac{4}{\sqrt{2\pi\sigma}}(x-\mu)\right\}\right]^{-1}, \quad (3)$$

where μ and σ^2 are the mean and variance, respectively which are calculated from the sequence used in section II.

Moreover, these four values are presented in parallel in this paper for simplicity.

(b) Sudden sound

There are many research studies regarding the detection of a specific sudden sound. For example, Mikami et al. [6] focused on a hitting sound for baseball by using intervals of the rise and decay of power, and Pikrakis et al. [7] focused on gunshots by using Bayesian networks. In the case of the detection of specific sounds, corresponding methods are needed. In this paper, we need to treat various sounds together.

We use the temporal difference of a short-term spectrum because a sudden sound sharply increases the power. The temporal difference of a short-term spectrum $s_b(i)$ for subband b and frame i (there are 5 subbands, frame length is 16msec, and frame shift is 50%) is calculated, and then the maximum value in a segment (1 second) is obtained.

However, using temporal differences may detect periodic sounds, such as the human voice. In order to avoid detecting such sounds, periodicity $bp(b)$ for each subband b and frame is also calculated. In order to underestimate the part with high periodicity, $1-|bp(b)|$ is used.

The product of the above-mentioned two values is calculated, and in order to normalize the value, the sigmoid function (3) is applied in the same manner as in the previous section.

B. *Audio information presentation on GUI*

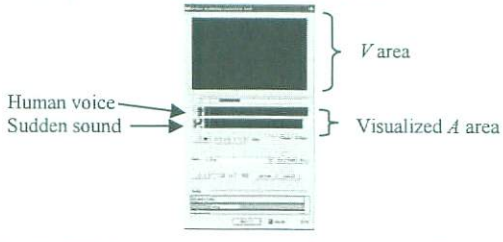

Fig. 4 GUI for evaluation (with A information visualization)

The A information is visualized as follows. Timelines for both "human voice" and "sudden sound" are displayed respectively as in Fig. 4. On each timeline, the value obtained in section IIIA is quantized by 256 steps, and then displayed as a marker in each segment (1 second).Therefore, the values are displayed within the range of 0 and 1. The subject extracts highlights while referring to this information in addition to V information.

IV. Experiments and Discussions

In order to evaluate the validity and usefulness of our proposed method, the following experiments are carried out.

A. *Experiment 1: Verification of audio feature detection*

In this section, the method proposed in section IIIA(a) is compared with audio indexing results, that is, ground truth obtained manually. The accuracy for an index l and an audio feature f is calculated as

$$(1/M)\sum_{t\in L}^{M} D(t,f) - (1/N)\sum_{t\in L}^{N} D(t,f), \quad (4)$$

where L is the set of segments labeled as an index l, $D(t,f)$ is the displayed value at segment t (the value is calculated by (3) and then quantized; $0 \le D(t,f) < 1$). M and N are the number of segments labeled and not labeled as an index l, respectively. The results of accuracy are shown in TABLE III.

From this table, it can be observed that the proposed method in section IIIA(a) can detect the human voice, in other words,

[1] The reference [4] also uses spectral information which is SNR for each subband. However, we do not use this feature because it is difficult to estimate the noise level from the original signal.

distinguish the "human voice" from others[2]. The amplitude information is very useful for non-noisy sound, but does not make sense for sound that includes loud noise. Even in this case, ZCR, GMM likelihood, and AI can detect the segments labeled "human voice." Actually, at least one of the feature values of ZCR, GMM-likelihood, and AI is taken more than 0.7 for "human voice" labeled segments. Moreover, the proposed method in section IIIA(b) can detect "sudden sound," such as machine noise and impulse noise.

TABLE III
MATCHING ACCURACY OF AUDIO FEATURE DETECTION

	Sec.IIIA(a) for human voice				Sec.IIIA(b) for sudden sound
	Amp.	ZCR	GMM	AI	
Human voice	0.26	0.41	0.44	0.39	0.05
Life sound	-0.01	0.06	-0.10	0.06	-0.08
Machine noise	-0.09	-0.06	-0.12	0.03	0.46
Impulse noise	-0.08	0.01	-0.17	-0.04	0.43

B. *Experiment 2: Verification of usefulness*

TABLE IV
MATCHING ACCURACY OF HIGHLIGHT EXTRACTION SEGMENTS

	# segments for *V+A*	Ins.	Del.	Co-corr.
Vonly	889	321	305	568(=63.9%)
V+visA		76	85	813(=91.5%)

In order to evaluate the usefulness of the proposed method, a subjective experiment, similar to that in section II, is conducted with the following exceptions (a) different test sequences are presented, and (b) the proposed method described in section III (hereafter denoted as "*V+visA*") is also executed. The presentation order is *Vonly*, *V+visA*, and *V+A*, so that the effects of audio visualization can be evaluated. The condition is shown in TABLE I.

The results are shown in TABLE IV. "Ins," "Del," and "Co-corr." are the comparison results for *V+A*, in the same manner as TABLE II. In other words, the degree of similarity is expressed for the extracted highlights of *Vonly* and *V+visA* compared to that of *V+A*.

This table shows that *V+visA* approaches that of *V+A* compared to *Vonly*. Due to application of the visualization of audio information, about 76.4% (=[0.915-0.639]/[1-0.639]) of the segments obtain the same results as that obtained when listening to the audio.

A typical improvement example caused by audio visualization is shown in Fig. 5. This figure shows the key frames, visualized audio data, and three kinds of extracted highlight information. A highlight is extracted from the last half for *Vonly*. On the other hand, the highlight is extracted from midway for *V+visA* because the "human voice" section is included there. This means that *visA* information leads *V+visA* to a similar result as *V+A*.

As an opinion from subjects, if the human voice segments are displayed, they want to know who is speaking (speaker recognition) and what is said (speech recognition). Since this paper focuses on a simple display method, they are outside the scope of this paper. However, if the utterances are displayed more simply and obviously, it is expected that the user interface would become more user-friendly.

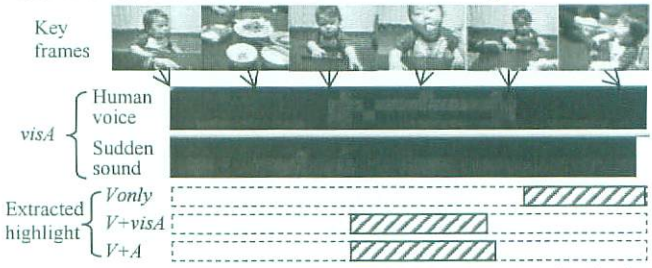

Fig. 5 A typical improvement example caused by *A* information visualization.

V. CONCLUSIONS

In this paper, in order to extract highlights from home video using only visual information on a mobile terminal, a method is proposed for visualization of audio information in the video.

First, the difference in the results of extracting highlights was examined between watching home video while listening to the audio and watching a silent home video. As a result, it was confirmed that "human voice" and "sudden sound" played important roles when watching a home video while listening to the sound.

Therefore, we tried to create a mechanism for home video users to efficiently extract highlights with visual information, including visualized audio information, by marking the segments corresponding to "human voice" and "sudden sound," which are detected automatically and presented on a timeline. It was confirmed that the results of extracting highlights only using the visualized information (without directly listening to the audio) were close to those using both video and audio.

In order to evaluate the contribution of each proposed method in section III, we need to evaluate the effects according to (a) whether audio information was properly visualized or not and (b) whether each audio feature was effectively applied or not. Since we have evaluated only the case (a), the case (b) needs to be evaluated in a future study.

REFERENCES

[1] K. Minami, A. Akutsu, H. Hamada, Y. Tonomura, "Video handling with music and speech detection," *J. IEEE Multimedia*, Vol. 5, Issue 3, pp. 17-25, 1998.

[2] Cees G.M. Snoek, Marcel Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, Vol. 25, 1, pp. 5-35, 2005.

[3] Truong, B. T. and Venkatesh, S., "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun.* Appl. 3, 1, Article 3, Feb. 2007.

[4] Yusuke Kida and Tatsuya Kawahara, "Evaluation of voice activity detection by combining multiple features with weight adaptation," *Proc. INTERSPEECH*, pp.1966-1969, 2006.

[5] H. Gustav Mueller and Mead C. Killion, "An easy Method For Calculating the Articulation Index," *Hearing Journal*, Vol. 43. No. 9, Sept. 1990.

[6] Dan Mikami, Seiichi Konya, Masashi Morimoto, "Pitch by Pitch Event Detection Using Impulse Sound Detection and Moving Image Clustering," *IEICE Trans. Info. and Sys.*, Vol.J90-D , No.2, pp.526-534, 2007.

[7] Aggelos Pikrakis and Sergios Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and Bayesian networks," *Proc. IEEE ICASSP2008*, pp.21-24, Mar. 2008.

[2] Each term in (4) actually ranged around from 0.2 to 0.8. For example, when ZCR was applied, the obtained average accuracy values for the segments labeled and not labeled as "human voice" were 0.72 and 0.31, respectively. Therefore, note that most of the obtained accuracy values, which were the difference of the two average values, did not become large.